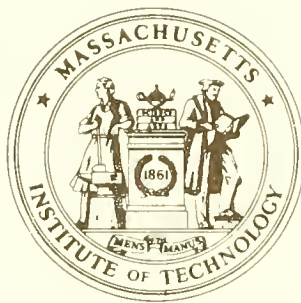


MIT LIBRARIES



3 9080 00441 5326



LIBRARY
OF THE
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

**working paper
department
of economics**

NON-RANDOM MISSING DATA

Jerry A. Hausman and A. Michael Spence*

Number 200

May 1977

**massachusetts
institute of
technology**

**50 memorial drive
cambridge, mass. 02139**



f

NON-RANDOM MISSING DATA

Jerry A. Hausman and A. Michael Spence^{*}

Number 200

May 1977

^{*} Massachusetts Institute of Technology and Harvard University, respectively. Research support from the National Science Foundation is gratefully acknowledged.



Digitized by the Internet Archive
in 2011 with funding from
Boston Library Consortium Member Libraries

<http://www.archive.org/details/nonrandommissing00haus>

NON-RANDOM MISSING DATA

by Jerry A. Hausman and A. Michael Spence¹

Missing data is an important problem in applied statistics and econometrics. It is well known that in a sample from a multivariate distribution if each observation is discarded for which one or more entries is missing, statistics computed from the reduced sample may not be appropriate estimates of their population counterparts. Likewise, in econometrics, estimation of structural models using only the "complete" data will lead to biased and inconsistent estimates of the true parameters if the realizations of the missing observations are correlated with the residuals in the equation being estimated. To overcome this problem, much work has been done to replace the missing entries with estimates derived from the available data. The simplest example of such procedures is to let the underlying distribution of a vector of random variables be multivariate normal and to assume that only one of the variables has missing observations. Unbiased estimates of the missing entries may be obtained by regressing the observed entries on the remaining data in each observation. The estimated regression parameters can then be used to form an unbiased prediction of the missing entries so long as the conditional expectation of the missing entries is the same as for the observed entries; that is, the probability of an observation being missing cannot depend on the value that would have been observed. Much more complicated situations where more than one variable may be missing in different patterns among the observations may occur, and these situations have been discussed by Afifi and Elashoff [1], Hartley and Hocking [7], Orchard and Woodbury [14],

¹We would like to thank G. Chamberlain, R. Gordon, Z. Griliches, Bronwyn Hall, R. Quandt, and D. Wise for helpful comments.

and recently by Dempster, Laird, and Rubin [4]. The techniques used are generalizations of the simple example in which the missing entries are conditioned on the available information and predictions are formed from the mean and variance of the multivariate normal distribution.

These procedures all depend on the conditional expectation for a variable remaining the same whether or not it is observed. Yet situations may occur where this assumption is invalid. Returning to the simple example, consider the case of a censored variable which is not recorded if its value is less than zero. The conditional expectation of the observed entries, conditional on both the remaining data in each observation and the fact that they are observed, exceeds the conditional expectation of the variable when the fact of censoring is not taken into account. Then predictions of the missing entries using estimated regression parameters from the observed entries will lead to upward biased predictions. Of course, in this simple example the problem would be readily apparent so that appropriate procedures could be used to form consistent predictions. However, a more subtle and less readily apparent problem may occur which can lead to similar consequences. Suppose the probability of observing the entry depends on its realization. That is, negative values of the variable have a higher probability of being missing but some negative values will still be observed. If simple regression techniques are used to form the predictions, biased and inconsistent predictions will again result. Thus a statistical procedure is required which will correct this problem if it is present and yield predictions with desirable properties.

In this paper we apply maximum likelihood procedures which can be used to test whether the problem exists and also to provide predictions of the missing entries. These procedures are closely related to work in the

econometric literature by Quandt [15], Goldfeld and Quandt [5], Hanoch [6], Hausman and Wise [8], Heckman [9], Lee [11], Maddala and Nelson [12], and Nelson [13]. Nelson has called this general problem the "stochastic censoring" problem. While our model differs from his specification, the problem being considered is similar. Along with the statistical specification of the problem we will discuss a method for finding the maximum likelihood estimates of the unknown parameters. Previous authors have reported difficulties in maximizing similar likelihood functions, but using a modified scoring method first employed by Berndt, Hall, Hall, and Hausman [8], excellent results are obtained. Alternative consistent but asymptotically inefficient estimators will also be considered in the appendix. Such estimators for related problems have been proposed by Amemiya [2], Hanoch [6], Heckman [9], and Lee [11]. Lastly, an example from a survey of 84 Canadian industries is used to demonstrate the usefulness of the procedure and the existence of the problem in actual data.¹

¹The problem came to the attention of the authors in the context of estimating a simultaneous equations model of the determinants of structure and performance in Canadian industry.

1. Statistical Specification

The most common specification for the missing data problem is to assume the observations z_i , $i = 1, \dots, T$, are drawn from a multivariate normal distribution $N(\mu, \Sigma)$ where z_i and μ are assumed to be vectors of length k and the covariance matrix Σ is an unknown $k \times k$ positive definite matrix. For the time being, we consider the case where some of the first entries are not observed; that is z_{1i} is missing for the incomplete observations. Reordering the observations so that the first s observations correspond to the complete observations, the sample is divided into two parts: $i = 1, \dots, s$ corresponds to all entries observed while $i = s+1, \dots, T$ corresponds to z_{1i} missing. Then using more traditional regression notation set z_{1i} equal to y_i and the $k-1$ subvector $(z_{2i}, \dots, z_{ki})$ equal to X_i so that

$$(1.1) \quad E(y_i | X_i) = X_i \beta = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (X_i - \mu_x)$$

$$V(y_i | X_i) = \sigma^2 = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$$

where μ_1 and Σ_{11} correspond to the mean and variance of z_1 while Σ_{ij} corresponds to the partitioning of z into y and X . To predict the missing $T-s$ values of y_i the following procedure is often used. For the sample $i = 1, \dots, s$ run least squares of y_i on X_i to estimate β , say $\hat{\beta}_{OLS}$. Then form the conditional predictions $\hat{y}_i = X_i \hat{\beta}_{OLS}$ for $i = s+1, \dots, T$. Under the assumption that $E(y_i | X_i, y_i \text{ not observed}) = E(y_i | X_i, y_i \text{ observed})$ then the predictions are known to be the best, linear unbiased predictions.

Problems arise, however, when the equality of the conditional expectations fails to hold. The simple censoring case corresponds to the Tobit model in econometrics (Tobin [16]). ^{In that case,} y_i is not observed if its realization is less than a constant, say zero. Then the conditional expectation of y_i

given X_i and the fact that it is observed is

$$(1.2) \quad E(y_i | X_i, y_i \geq 0) = X_i \beta + \sigma \frac{\phi(X_i \beta / \sigma)}{\Phi(X_i \beta / \sigma)}$$

where ϕ and Φ correspond to the unit normal density and distribution function, respectively.¹ This result follows from writing $y_i = X_i \beta + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$ and noting that y_i is observed if $\varepsilon_i \geq -X_i \beta$. The expected value of ε_i then follows from

$$(1.3) \quad \begin{aligned} E(\varepsilon_i | X_i, \varepsilon_i \geq -X_i \beta) &= \frac{1}{1 - \Phi(-X_i \beta / \sigma)} \int_{-X_i \beta}^{\infty} \frac{\varepsilon_i}{\sqrt{2\pi\sigma^2}} e^{-\varepsilon_i^2 / 2\sigma^2} d\varepsilon_i \\ &= \frac{1}{1 - \Phi(-X_i \beta / \sigma)} \int_{-X_i \beta / \sigma}^{\infty} r \phi(r) dr = \sigma \frac{\phi(X_i \beta / \sigma)}{\Phi(X_i \beta / \sigma)} \end{aligned}$$

On the other hand, for those y_i 's which are not observed

$$(1.4) \quad E(y_i | X_i, y_i < 0) = X_i \beta - \sigma \frac{\phi(X_i \beta / \sigma)}{1 - \Phi(X_i \beta / \sigma)}$$

If regression coefficients formed from the good data corresponding to equation (1.2) are used to predict the missing data corresponding to equation (1.4), an upward bias results. In this situation estimates of β and σ may be estimated from the log likelihood function (Tobin [17] and Amemiya [2]).²

$$(1.5) \quad \ell = k - \frac{S}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^S (y_i - X_i \beta)^2 + \sum_{i=S+1}^T \log [1 - \Phi(X_i \beta / \sigma)]$$

Then consistent conditional expectations would be formed using equation (1.4). Alternative estimators [which require less computation] than maximum likelihood are available. They will be discussed in the appendix. However, numerous algorithms and computer programs exist which yield maximum likelihood estimates at relatively low cost even for very large samples.

¹The term $\phi(X_i \beta / \sigma) / \Phi(X_i \beta / \sigma)$ corresponds to the inverse Mills' ratio. The denominator gives the probability that y_i is observed conditional on X_i .

²The constant term k in the log likelihood function equals $-\frac{S}{2} \log (2\pi)$.

The particular case of censoring is apt to be evident in a sampling situation. A more common but less evident sampling situation arises when the probability of observing y_i varies with its realization and perhaps the realizations of other random variables W_i . Then defining the indicator $d_i = 1$ for y_i being observed and $d_i = 0$ if y_i is not observed we might specify a model where the probability of observing y_i depends on its realization and other variables. First, write y_i in a regression equation as $y_i = X_i\beta + \varepsilon_{1i}$. A general linear specification could make the probability of observing y_i take the following form: y_i is observed if $\alpha y_i + X_i\theta + W_i\gamma + \eta_i \geq 0$ where W_i are variables which do not enter the conditional expectation but affect the probability of y_i being observed and the η_i are iid random variables. Substituting for y_i leads to $X_i(\alpha\beta + \theta) + W_i\gamma + \alpha\varepsilon_{1i} + \eta_i$. But since α and θ enter the specification in an equivalent way it is convenient to combine them into a vector and write the specification as $X_i\xi + W_i\gamma + \varepsilon_{2i}$ which corresponds to the "reduced form" of the model where $\varepsilon_{2i} = \alpha\varepsilon_{1i} + \eta_i$. Defining the vectors $R_i = [X_i \ W_i]$ and $\delta = \begin{bmatrix} \xi \\ \gamma \end{bmatrix}$ then gives a probability of y_i being observed

$$(1.6) \quad \text{pr}(d_i = 1) = \text{pr}(R_i\delta + \varepsilon_{2i} \geq 0)$$

Then assuming that η_i is normally distributed and taking the variance of ε_2 equal to 1, $\sigma_2^2 = 1$, as a normalization, we have the probit model

$$(1.7) \quad \text{pr}(d_i = 1) = \Phi(R_i\delta) \quad \text{and} \quad \text{pr}(d_i = 0) = 1 - \Phi(R_i\delta).$$

However, besides knowing whether $d_i = 0$ or $d_i = 1$, we also know the value of y_i if it is observed, $d_i = 1$. This observation yields information about ε_{1i} which in turn gives information about ε_{2i} if they are correlated. Thus in specifying the likelihood for those observations with $d_i = 1$, equation (1.7) will be altered to use this additional knowledge.

One of two sample outcomes occur for each observation. The first outcome is $d_i = 1$ so that y_i is observed along with X_i and R_i . Writing the joint (generalized) density and ^{then} conditioning on the observed value of y_i gives the expression

$$(1.8) \quad f(y_i, d_i = 1 | X_i, R_i) = g(d_i = 1 | y_i, X_i, R_i) h(y_i | X_i) \\ = \text{pr}(R_i \delta + \varepsilon_{2i} \geq 0 | y_i, X_i, R_i) \frac{1}{\sigma_1} \phi \left(\frac{y_i - X_i \beta}{\sigma_1} \right).$$

Then since ε_{1i} and ε_{2i} are bivariate normal ^{with correlation ρ} $\wedge$ the regression equation for ε_{2i} given ε_{1i} is $\varepsilon_{2i} = \frac{\rho}{\sigma_1} \varepsilon_{1i} + v_i$ where v_i is $N(0, 1 - \rho^2)$ and is uncorrelated with ε_{1i} using $\sigma_2 = 1$ from the earlier normalization. Using this relationship equation (1.8) may be rewritten as

$$(1.9) \quad f(y_i, d_i = 1 | X_i, R_i) = \text{pr}(R_i \delta + \frac{\rho}{\sigma_1} (y_i - X_i \beta) + v_i \geq 0) \frac{1}{\sigma_1} \phi \left(\frac{y_i - X_i \beta}{\sigma_1} \right) \\ = \phi \left(\frac{R_i \delta + \frac{\rho}{\sigma_1} y_i - \frac{\rho}{\sigma_1} X_i \beta}{(1 - \rho^2)^{1/2}} \right) \frac{1}{\sigma_1} \phi \left(\frac{y_i - X_i \beta}{\sigma_1} \right).$$

Having derived the likelihood when y_i is observed, the case where y_i is not observed so that $d_i = 0$ follows easily. Basically y_i is "integrated out" of equation (1.8) to find the likelihood as expected from equation (1.6)

$$(1.10) \quad f(d_i = 0 | X_i, R_i) = \text{pr}(R_i \delta + \varepsilon_{2i} \leq 0 | X_i, R_i) \\ = 1 - \phi(R_i \delta).$$

Using the ordering of the data where the first s observations correspond to y_i being observed and the last $T - s$ values of y_i being missing, the

log likelihood function follows from equations (1.9) and (1.10). Setting c , a constant, equal to $-\frac{s}{2} \log 2\pi$ yields the log likelihood function

$$(1.11) \quad \ell = c - \frac{s}{2} \log \sigma_1^2 + \sum_{i=1}^s \left[\log \Phi \left(\frac{R_1 \delta + \frac{\rho}{\sigma_1} (y_i - X_i \beta)}{(1 - \rho^2)^{\frac{1}{2}}} \right) - \frac{1}{2\sigma_1^2} (y_i - X_i \beta)^2 \right] \\ + \sum_{i=s+1}^T \log (1 - \Phi(R_1 \delta)).$$

Maximization of the log likelihood function in equation (1.11) leads to estimates of $\theta = (\beta, \delta, \rho, \sigma_1)$ which have the usual desirable properties of maximum likelihood estimates. Although this specification does not satisfy the classical regularity conditions since the likelihood is a mixture of densities and probabilities, Amemiya's [1] proofs for the Tobit case carry over in a straightforward manner for this problem.

The crucial role of the covariance between ε_{1i} and ε_{2i} is evident in the correlation coefficient ρ in equation (1.9).¹ If $\rho = 0$ so that no correlation exists, then the probability of observing y_i is independent of ε_{1i} and the least squares procedure of predicting the missing y_i 's from the regression of the observed y_i 's is a correct procedure. On the other hand, if $\rho \neq 0$ then the probability of observing y_i depends on ε_{1i} so that the regression procedure will lead to the incorrect conditional expectation. The conditional expectation is calculated from again writing $y_i = X_i \beta + \varepsilon_{1i}$ and noting that the regression equation for ε_{1i} given ε_{2i} may be written as $\varepsilon_{1i} = \rho \sigma_1 \varepsilon_{2i} + \omega_i$ where $\omega_i \sim N(0, \sigma_1^2(1 - \rho^2))$ and is independent of ε_{2i} (the normalization $\sigma_2^2 = 1$ has again been used). Then the conditional expectation of ε_{1i} given that $d_i = 1$ or equivalently that $\varepsilon_{2i} \geq -R_1 \delta$ is

¹ Since $\varepsilon_{2i} = \alpha \varepsilon_{1i} + \eta_i$, $\rho \neq 0$ when $\alpha \neq 0$ which is equivalent to having the probability of observing y_i depend on its realization.

$$\begin{aligned}
 (1.12) \quad E(\varepsilon_{1i} | \varepsilon_{2i} \geq -R_i \delta) &= \rho \sigma_1 E(\varepsilon_{2i} | \varepsilon_{2i} \geq -R_i \delta) + E(\omega_i | \varepsilon_{2i} \geq -R_i \delta) \\
 &= \rho \sigma_1 \frac{\phi(R_i \delta)}{\Phi(R_i \delta)}
 \end{aligned}$$

using equation (1.3) and the normalization of $\sigma_2^2 = 1$. Then the conditional expectation of y_i follows

$$(1.13) \quad E(y_i | X_i, d_i = 1) = X_i \beta + \rho \sigma_1 \frac{\phi(R_i \delta)}{\Phi(R_i \delta)} .$$

Thus, as in the earlier case, least squares on the observed data leads to biased and inconsistent estimates of β . Furthermore, predictions for the missing data will be inconsistent since

$$(1.14) \quad E(y_i | X_i, d_i = 0) = X_i \beta - \rho \sigma_1 \frac{\phi(R_i \delta)}{1 - \Phi(R_i \delta)} .$$

To form consistent predictions, consistent estimates of the parameters β, δ, ρ , and σ_1 are required, and equation (1.14) is used to form the expectations.

Unfortunately, the log likelihood function does not appear to be globally concave so that some care must be taken in choosing an algorithm to find the maximum. Furthermore, some authors have reported difficulty in broadly similar situations. Therefore, the next section discusses an algorithm which has performed quite well in practice.

2. Estimation

To maximize the likelihood function from equation (1.11), an algorithm proposed by Berndt, Hall, Hall, and Hausman [3] is used. This method is similar to the method of scoring, e.g. Rao [16, pp. 366ff], but instead of requiring large sample expectations to be taken to evaluate Fisher's information matrix, it relies on the result that in large samples the covariance of the gradient is a consistent estimate of the information matrix in the neighborhood of the optimum. Letting θ be the parameter vector and the gradient $g(\hat{\theta}) = \frac{\partial \ell}{\partial \theta} \big|_{\hat{\theta}}$, its asymptotic covariance in the neighborhood of the optimum is approximated by $Q(\hat{\theta}) = \sum_{i=1}^T \frac{\partial f_i}{\partial \theta} \big|_{\hat{\theta}} \frac{\partial f_i}{\partial \theta} \big|_{\hat{\theta}}'$ where f_i is the $\log$ likelihood of the i th observation. Thus, the method requires only computation of first derivatives. The updating formula at the j th iteration for $\hat{\theta}^{j+1}$ is then

$$(2.1) \quad \hat{\theta}^{j+1} = \hat{\theta}^j + \lambda^j Q(\hat{\theta}^j)^{-1} g(\hat{\theta}^j)$$

where $\lambda^j > 0$ is the stepsize chosen in the direction $Q(\hat{\theta}^j)^{-1} g(\hat{\theta}^j)$. The stepsize λ^j is chosen according to the criterion in Berndt, et al., p. 656, and convergence to a local maximum is assured. The algorithm has the desirable "uphill" property: an improvement in the likelihood function occurs at each step. Experience with this algorithm has been very satisfactory so long as the derivatives are calculated correctly.¹ The algorithm can be considered a generalization of the Gauss-Newton algorithm. Estimates of the asymptotic covariance matrix of the estimates follow from $Q(\hat{\theta}_{ML})^{-1}$, where $\hat{\theta}_{ML}$ is the value of θ which maximizes the likelihood function. Care must be taken to insure that the global maximum has been found although in practice no difficulties arose.

¹This statement is not meant to imply that numerical derivatives are inferior to analytical calculation of derivatives by the computer. Rather, all convergence problems to date have been caused by incorrect algebraic computations of the derivatives.

Although initial consistent estimates may be derived from methods discussed in the appendix and used as starting values for the maximum likelihood algorithm, an easier procedure performed well in practice. Least squares estimates on the observed data are used to set initial values for $\hat{\beta}^1$ and $\hat{\sigma}_1^1$ while $\hat{\delta}^1 = 0$ and $\hat{\rho}^1 = 0$. The maximum likelihood algorithm always converged to the global optimum beginning from these estimates. This maximum likelihood seems to provide a relatively inexpensive estimator with little trouble encountered in computing the estimates.¹

¹On the sample of 84 observations discussed in the next section, computer costs on the 370-168 at MIT averaged about \$3. For a larger problem with nearly 2000 observations, the cost increased to about \$8.50.

3. Empirical Example

We now apply the statistical model specified in Section 1 to a body of data collected on the economic characteristics of Canadian and US industries. The sample consists of data on 84 industries with the number of missing observations varying from zero to 35. While many of the characteristics are continuous and appear approximately normal after transformation, other characteristics are ^{represented} by dummy variables. These dummy variables might well be expected to enter the vector R_1 in equation (1.6) even if they did not appear in the conditional mean. Two variables with missing observations were chosen to be studied. The first variable, FSE, represents the amount of ^{Canadian} foreign control in the industry. It is constructed by dividing the value of shipments by establishments classified as 50% or more foreign-controlled by the value of shipments by all establishments in the industry. It was felt that the missing observations here might well correspond to low values of FSE since data are not collected on foreign ownership unless it is "significant." Twenty-four out of 84 industries are missing observations on FSE. The second variable studied, CDRC, represents the cost disadvantage of small firms. It is constructed by dividing the value added per worker in the smallest one-half firms in an industry by the value added per worker in the other one-half of the industry. For CDRC, 13 out of 84 industries have missing observations. However, no definite relation between a missing observations and the value of the variable was thought to exist.

To test for "biased" missing data, a test of the hypothesis $\rho = 0$ in equation (1.9) is undertaken. Inspection of that equation or the likelihood function in equation (1.11) demonstrates that if $\rho = 0$, the likelihood function factors into two parts: the probit part representing the probability

of observing the variable and a least squares part representing the density of the variable. Thus, the hypothesis can be tested by doing a likelihood ratio test comparing the maximum of equation (1.11) with the sum of the probit and least square likelihoods when $\rho = 0$. An asymptotically equivalent test used here which is computationally easier is to do a χ^2_1 test using the computed asymptotic variance of $\hat{\rho}$. The test statistic is computed by dividing $\hat{\rho}^2$ by the estimate of its asymptotic variance. Besides testing this hypothesis, we also compare the estimated $\hat{\beta}_{ML}$ to the least squares $\hat{\beta}_{OLS}$. Lastly, we form the conditional predictions for the missing data using equation (2.3) and compare them to the predictions using the least squares estimates, $\hat{y}_i = X_i \hat{\beta}_{OLS}$. This last comparison captures the most important aspect of the problem, whether multivariate normal methods (least squares) provide consistent predictions of the missing parameters.

In specifying the model for the extent of foreign ownership, FSE, the following variables are used to form the conditional mean: a constant, value added per establishment in the industry (VPE), the standard deviation of the value of shipments of firms in the industry (SSI), proportion of shipments in the US counterpart industry by the largest four enterprises in that industry (US467), the proportion on nonproduction workers in the industry (NPW), and average overhead labor costs in the industry (WNP). All these variables were included in the probit function of equation (1.6) along with estimates of a dummy variable whose value equals one for consumer goods industries if they sell a convenience good and is zero otherwise (CON). In Table 1 the maximum likelihood estimates of the parameters are presented along with their asymptotic standard errors. We also present the least squares estimates of the same model. Note that while the maximum likelihood estimates (MLE) have the same sign and order of magnitude as the least

TABLE I

Estimates for Extent of Foreign Ownership (FSE)

<u>Variable</u>	<u>Coefficient</u>	<u>MLE</u>	<u>OLS</u>	<u>MLE/OLS</u>
Constant	β_0	-.972 (.308)	-.696 (.242)	1.40
VPE	β_1	-.078 (.038)	-.092 (.044)	.85
SSI	β_2	-.970 (.943)	-1.13 (.576)	.85
US467	β_3	.077 (.022)	.097 (.016)	.79
NPW	β_4	.704 (.193)	.499 (.149)	1.41
WNP	β_5	.144 (.046)	.118 (.036)	1.22
Constant	δ_0	1.11 (1.57)		
VPE	δ_1	.099 (.051)		
SSI	δ_2	-.257 (.367)		
US467	δ_3	-.027 (.010)		
NPW	δ_4	.153 (.181)		
WNP	δ_5	.033 (.245)		
CON	δ_6	-.633 (.325)		
	ρ	.937 (.094)		
ℓ^* (log likelihood value)		-160.9		

squares estimates (OLS), notable differences in the coefficient estimates are present with the largest difference being about 40%. Examination of the estimate of ρ shows that the value of FSE significantly affects the probability of it being observed. The estimate of .937 exceeds its estimated asymptotic standard error by a factor of almost 10 and the asymptotic χ^2_1 test that $\rho = 0$ has the value of 99.4 which leads to a decisive rejection of the null hypothesis. For comparison purposes a likelihood ratio test was done on the ML value versus the sum of the likelihoods of a probit estimation and the least squares regression. A test of $\rho = 0$ follows from minus twice the difference of the log likelihood which equals 7.90 and is distributed as χ^2_1 . Again we reject the null hypothesis although not as decisively as before. Also $\hat{\rho}$ has the expected sign since examination of equation (1.9) shows that large values of the dependent variable, FSE, increase the probability of observing it, $d_i = 1$, for ρ taking positive values. In fact, this effect is so strong that the situation is almost the pure censoring case of equation (1.3) which has $\rho = 1$ when the censoring point is known. This result is in accord with the prior judgment of the economists who gathered the data that low values of FSE are less likely to be recorded in the data. Lastly, the mean of the 24 missing observations is estimated using equation (1.14) and the maximum likelihood estimates. Its value is .2195 which is considerably lower than the least squares prediction of .5724. Thus, the maximum likelihood estimates seem valuable in finding significant non-randomness in the missing data.

For comparison purposes, the technique is applied to the cost disadvantage ratio (CDRC) which has 13 out of 84 industries missing observations. Here, it was felt that the value of the variable itself might not significantly affect the probability of it being observed. The following variables are used to form the conditional mean: a constant, the standard deviation of the total value of shipments (SSI), average assets per 

employee (LAB2C), and the proportion of non-production workers in the industry (NPW). An additional variable, the proportion of shipments by the US counterpart industry by the largest form enterprises (US467), is used in the probit function. The results presented in Table II show that the ML estimates of the parameters of the conditional mean are extremely close to the least squares estimates. While $\hat{\rho}$ indicates some probability of observations of CDRC being missing for high values of CDRC, it is estimated quite imprecisely. It is less in absolute value than its asymptotic standard error and the χ^2_1 test for $\rho=0$ has a value of only .5. A fact which is perhaps more important is that the estimated mean for the 13 missing observations using the maximum likelihood estimates is 1.10 compared to the mean of the least squares estimates of .941. These predictions are quite close for a variable which has a standard deviation of 1.71 in the observed data. Thus, little evidence has been discovered that the missing observations do not form a random sample conditional on the X_i 's. Furthermore, the predicted values of the missing observations confirm the prior judgment of the economists who constructed the data.

In this section we have applied the missing data specification to two different variables. The results are in accord with prior judgment. Also, the maximum likelihood technique performed well in finding a potential bias in the missing observations and had a large effect in predicting the values of the missing observations.

TABLE II

Estimates of Cost Disadvantage Ratio (CDRC)

<u>Variable</u>	<u>Coefficient</u>	<u>MLE</u>	<u>OLS</u>	<u>MLE/OLS</u>
Constant	β_0	1.25 (.087)	1.24 (.087)	1.01
SSI	β_1	-1.24 (.629)	-1.28 (.552)	.97
LAB2C	β_2	-.103 (.066)	-.113 (.063)	.91
NPW	β_3	-.408 (.143)	-.407 (.150)	1.00
Constant	δ_0	2.89 (1.68)		
SSI	δ_1	-2.10 (4.47)		
LAB2C	δ_2	-.708 (.703)		
NPW	δ_3	-.485 (1.89)		
U5467	δ_4	-.262 (.143)		
	ρ	-.492 (.668)		
ℓ^* (log likelihood value)		-14.1		

4. Conclusion

In this paper a method has been proposed for predicting the values of missing data when the probability of observing the variable depends on its value. Basically, a univariate approach has been taken by concentrating on one variable at a time. Extension to the multivariate case where more than one variable may be missing seems straightforward, but may be complicated to do computationally. An equation of the form of equation (1.6) could be specified for each variable and then missing values of R_i replaced by their expected values. But since the variance of ε_{2i} then depends on the prediction variance of the missing values of R_i which would differ across observations, a fairly complex likelihood function would result with the probit function varying with each pattern of missing observations. However, the increased complexity of this situation might be worthwhile given the increased efficiency of using multivariate procedures.

References

- [1] Afifi, A. A. and R. M. Elashoff, "Missing Observations in Multivariate Statistics," Journal of the American Statistical Society, 62 (1967), 10-29.
- [2] Amemiya, T., "Regression Analysis when the Dependent Variable is Truncated Normal," Econometrica, 41 (1973), 997-1016.
- [3] Berndt, E., B. Hall, R. Hall, and J. Hausman, "Estimation and Inference in Nonlinear Structural Models," Annals of Economic and Social Measurement, 3 (1974), 653-65.
- [4] Dempster, A. P., N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, 1977, forthcoming.
- [5] Goldfeld, S. M. and R. E. Quandt, "The Estimation of Structural Shifts by Switching Regressions," Annals of Economic and Social Measurement, 2 (1973), 475-85.
- [6] Hanoch, G., "A Multivariate Model of Labor Supply: Methodology for Estimation," September 1976, mimeo.
- [7] Hartley, H. O. and R. R. Hocking, "The Analysis of Incomplete Data," Biometrics, 27 (1971), 783-808.
- [8] Hausman, J. A. and D. A. Wise, "Attrition and Sample Selection in the Gary Income Maintenance Experiment," 1971, mimeo.
- [9] Heckman, J., "Shadow Prices, Market Wage, and Labor Supply," Econometrica, 42 (1974), 679-94.
- [10] _____, "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement, 5 (1976), 475-92.

- [11] Lee, L. F., "Estimation of Some Limited Dependent Variable Models by Two Stage Methods," 1975, mimeo.
- [12] Maddala, G. S. and F. D. Nelson, "Switching Regression Models with Exogenous and Endogenous Switching," Proceedings of the Business and Economics Statistics Section, American Statistical Association, 1975, 423-26.
- [13] Nelson, F. D., "Censored Regression Models with Unobserved, Stochastic Censoring Thresholds," 1976, mimeo.
- [14] Orchard, T. and M. A. Woodbury, "A Missing Information Principle: Theory and Applications," Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1972.
- [15] Quandt, R. E., "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," Journal of the American Statistical Association, 53 (1958), 873-80.
- [16] Rao, C. R., Linear Statistical Inference and its Application, New York: John Wiley & Sons, Inc., 1973 (second edition).
- [17] Tobin, J., "Estimation of Relationship for Limited Dependent Variables," Econometrica, 26 (1958), 24-36.

Appendix

Alternative consistent, but inefficient, methods of estimation which may require less computation are also available. Hanoch [6] Heckman [9], and Lee [11] have used the conditional expectation of the observed data in equation (1.13) to derive consistent estimates of the structural parameters. An estimate of the inverse Mills ratio, $M_i = \phi(R_i\delta)/\Phi(R_i\delta)$ is made by using a maximum likelihood probit program to estimate δ . Then a regression can be run on the observed data to estimate β and σ_{12}

$$(2.2) \quad y_i = X_i\beta + \sigma_{12}\hat{M}_i + u_i \quad i = 1, \dots, s$$

Lastly, the consistent estimates of β and σ_{12} can be used to predict the missing y_i for $i = s+1, \dots, T$ using equation (1.14). However, to correctly do inference on the parameters, say $\sigma_{12} = 0$, an additional measure must be taken since u_i is heteroscedastic with $\text{var}(u_i) = \sigma_1(1 + \phi(R_i\delta)M_i - M_i)^2$. A weighted least squares procedure using estimates from the probit estimates can be used here. For the data used in this paper this procedure was not as satisfactory as maximum likelihood. The expense of doing both a probit estimation and then a least squares regression was only slightly less than the expense of doing maximum likelihood estimation. Also, for our sample of 84 observations the estimates differed markedly from the maximum likelihood estimates in some cases, although this criticism must be tempered with the knowledge that the true parameters are unknown so that no assurance exists that maximum likelihood is providing better estimates. However, by using only the conditional expectation this regression method seems likely to yield imprecise estimates unless the sample is large. The problem arises because the inverse Mills ratio has high correlation with X_i except for extreme probabilities.

For example, in the first case of the extent of foreign control this procedure estimated ρ to be only .03 with a standard error of .84, and the estimates of β are quite close to the least squares estimates. Therefore, its estimate of the mean of the missing observations is .5681 only slightly below the least squares prediction of .5724 and far above the maximum likelihood estimate of .2195. In the second example of the cost disadvantage ratio, its mean of the missing observations is 1.03 which is about halfway between the least squares estimate of .94 and the maximum likelihood estimate of 1.10. Thus, this consistent but inefficient procedure did not perform well in our example of 84 observations.

This combined probit-least squares procedure, however, may have important uses in large data sets where its consistency property insures accurate estimates. Also, beginning with these consistent estimates, one step of the maximum likelihood algorithm setting $\lambda^1 = 1$ in equation (2.1) yields estimates with the same asymptotic distribution as the maximum likelihood estimates as pointed out in Berndt, et al., [3 , p. 659]. However, given the low cost of computing the global maximum to the likelihood function, this one step efficient algorithm was not used. One last estimator should be mentioned. Amemiya [2] , has proposed a consistent, but inefficient, instrumental variables estimator for the Tobit specification which may be readily adapted to the current problem. It requires considerably less computation than the combined probit-regression estimator. However, it was not used here since it did not provide a convenient test of the hypothesis of randomness, $\rho = 0$, which we wanted to test in the current problem.



Date Due

OCT 26 '78

MAR 09 1998

102F

MAY 17

SEP 06

APR 27 1998

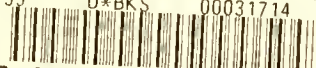
JUN 16 2004

T-J5 E19 w no.191
Krugman, Paul./Contractionary effects
730156 D*BKS 00031715



3 9080 000 795 440

T-J5 E19 w no.192
Diamond, Peter/Welfare analysis of imp
730159 D*BKS 00031714



3 9080 000 795 424

T-J5 E19 w no.193
Weitzman, Mart./Optimal revenue functio
730161 D*BKS 00031713



3 9080 000 795 408

HB31.M415 no.194
Fisher, Frankl/The effect of simple sp
731574 D*BKS 00037733



3 9080 000 867 090

HB31.M415 no.195
Diamond, Peter/Tax incidence in a two
731576 D*BKS 00037734



3 9080 000 867 116

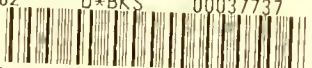


3 9080 004 415 300



3 9080 004 415 318

HB31.M415 no.198
Dixit, Avinash/Quality and quantity co
731582 D*BKS 00037737



3 9080 000 867 173

HB31.M415 no.199
Friedlaender, /Capital taxation in a d
731584 D*BKS 00037738



3 9080 000 867 199



3 9080 004 415 326

